View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Elsevier - Publisher Connector

JOURNAL OF MULTIVARIATE ANALYSIS 10, 611–627 (1980)

Consistent Nonparametric Regression from Recursive Partitioning Schemes*

LOUIS GORDON AND RICHARD A. OLSHEN†

Energy Information Administration, United States Department of Energy, Washington, D.C. 20461, and University of California, San Diego, La Jolla, California 92093

Communicated by M. Rosenblatt

We here extend our results on asymptotically Bayes risk efficient classification to the general regression scenario. More precisely, we find L^p consistent estimators for an arbitrary regression function provided only that the dependent variable has a finite absolute pth moment. The estimators are truncated and untruncated local means derived from recursive partitioning schemes.

1. INTRODUCTION AND SUMMARY

In this paper certain estimators for the general regression problem are shown to converge in pth mean, $p \geqslant 1$, to the true regression function whenever the latter possesses an absolute pth moment. The estimators are truncated, and untruncated, local sample means which flow from recursive partitioning schemes. The results extend our previous work on the Bayes risk consistency of rules for the classification problem [6].

Suppose that $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed (iid), $X \in R^d$, $Y \in R^1$. If f is a function on R^1 for which $E\{f(Y)\}$ exists, then $h(x) = E\{f(Y) \mid X = x\}$ is the regression of $f(Y)$ on X. The assumption is that X, $\{(X_i, Y_i)\}_{i=1}^n$, and f are given and that h is to be estimated. For convenience, we have taken f to be the identity function. The case for which Y assumes only two values is an analogue of the "linear logistic" problem for binary data (see Cox [3]). If $\hat{h}_n$, our estimator, is required also to take only two values, the problem is a special case of the (two population) "classification" problem (see [6]).

Received November 25, 1978; revised May 2, 1980.

AMS 1970 subject classifications: Primary 62G05; Secondary 62E20.

Key words and phrases: Nonparametric regression, recursive partitioning schemes, convergence in pth mean.

* Research supported in part by National Science Foundation Grants MCS76-08314 and MCS-7906228 to the University of California, San Diego.

† Completed in part at the Massachusetts Institute of Technology and Harvard University.

0047-259X/80/040611-17$02.00/0

Copyright © 1980 by Academic Press, Inc.
All rights of reproduction in any form reserved.

In our last paper we demonstrated that variations of rules introduced (separately) by Friedman [5], by Anderson (statistically equivalent blocks—see [1]), and by Morgan and Sonquist (AID—see [8]) lead to asymptotically Bayes risk efficient solutions to the two-population classification problem no matter what the conditional distributions of X given Y are. Trivial extensions of our arguments yield the same general conclusions for the k-population classification problem.

All the rules cited in the previous paragraph are invariant under strictly monotonic transformations of the X coordinate axes. Likewise, the corresponding rules for the regression problem studied here are invariant, an invariance shared by the conditional expectation itself. Thus, without loss, we assume that $X \in U$, the unit cube in R^d. As stated, we assume that $E\{|Y|^p\} < \infty$ for some $p \geqslant 1$. Our theorem shows that for a large class of estimators $\hat{h}_n$ similar to those discussed in [6], $E\{|\hat{h}_n - E(Y|X)|^p\}$ tends to 0 as n tends to infinity. These estimators are of the following general form. Based on $(X_1, Y_1), \ldots, (X_n, Y_n)$, U is recursively partitioned into "boxes." For each box the average—or possibly a truncated average—of the Y's within that box is computed. This average is the value of $\hat{h}_n$ for any X which falls into the box. (In two-population classification, a relative majority vote is taken with the "winner" the value of $\hat{h}_n$ within the box; see [6].)

Informally, a box is a rectangular parallelepiped with sides parallel to the coordinate axes, together with certain parallelepipeds which are subsets of the sides. In order to make this paper reasonably self-contained we include in the next section formal notation and definitions for boxes and partitions. The reader will note that a recursive partitioning scheme has associated with it a binary tree. See [2, 5] for the details of this association.

The results of this paper are for the same statistics problem as the one discussed by Stone [10], by Devroye and Wagner [4], and by Spiegelman and Sacks [9]. In the first, Stone gives necessary and sufficient conditions for convergence in pth mean of nearest-neighbor estimators of a general regression function. In the latter two, Devroye and Wagner and Spiegelman and Sacks give sufficient conditions for convergence in pth mean of kernel and window estimators of a general regression function. Our estimators are quite different from those of any of the previous three papers. Our principal argument is quite different, also. We argue that the sequence of conditional expectations of an L^p function given the cited partitions tends in L^p to the function as certain norms of the partitions tend to 0. Moreover, a large deviation result of Kiefer [7] is employed several times. In our closing section we show how Stone's work can be brought to bear on the estimators we study, and in some detail we apply Stone's approach to Anderson's rules [1] for general nonlinear regression.

While we do not give a detailed listing of estimators to which the results of the present paper apply, we do note that: (1) several explicit algorithms

for the classification problem are discussed in [6]; (2) the AID algorithms [8] are widely used in the social sciences, and our theorems apply to variations of them; (3) a monograph on recursive partitioning rules for classification and regression by Breiman *et al.* is in preparation [2], and our theorems apply to some of the rules in that monograph; (4) Friedman's program is finding increasing use in the medical field (see [11]).

2. Boxes and Partitions

The preceding section contains an informal discussion of boxes and partitions. However, more precise definitions are necessary to an understanding of the remainder of the paper. Therefore, this section gives those definitions from Section 2 of [6] which we require in subsequent work.

A *basic box* in R^d is a triple (a, b, r) of vectors, $a, b \in R^d$, and $r_i \in \{0, 1, 2, 3\}$ for $i = 1, 2, \ldots, d$. We identify a basic box with the subset

$$\begin{aligned} B = & \bigcap_{\{r_i=0\}} \{x \in R^d \mid a_i < x_i < b_i\} \\ & \cap \bigcap_{\{r_i=1\}} \{x \in R^d \mid a_i \leqslant x_i < b_i\} \\ & \cap \bigcap_{\{r_i=2\}} \{x \in R^d \mid a_i < x_i \leqslant b_i\} \\ & \cap \bigcap_{\{r_i=3\}} \{x \in R^d \mid a_i \leqslant x_i \leqslant b_i\}. \end{aligned} \tag{2.1}$$

A *vertex* of (a, b, r) is a vector v in R^d with $v_i = a_i$ or $v_i = b_i$ for all dimensions i. The vertex b is called the *upper vertex* and a is the *lower vertex* of the basic box (a, b, r).

The *dimension* of (a, b, r) is the cardinality of $\{i \mid a_i < b_i\}$.

A *subside* of a basic box $B = (a, b, r)$ is a basic box $B' = (a', b', r')$ for which:

(i) there exists a dimension i_0 for which $a_{i_0} < b_{i_0}$ and $a'_{i_0} = b'_{i_0} = a_{i_0}$ or $a'_{i_0} = b'_{i_0} = b_{i_0}$;

(ii) $a_i \leqslant a'_i \leqslant b'_i \leqslant b_i$ for all dimensions i;

(iii) at least one vertex of B' is a vertex of B.

A box is a union of a basic box and a set of subsides such that, for each dimension i, for at most one subside is $a'_i = b'_i = b_i$ and for at most one subside is $a'_i = b'_i = a_i$. Note that by definition, any box may be considered a union of at most $2d + 1$ basic boxes, and that all basic boxes are convex.

We reserve Q as a generic symbol for a finite partition of U, all of whose

component subsets are boxes B. For $x \in U$, we denote by $B(x)$ the unique box in Q containing x. The upper and lower vertices of B are denoted $b(B)$ and $a(B)$. We occasionally suppress the explicitly stated dependence of a and b on B. If a sequence of partitions is discussed, the index is superscripted, and the same indexing is carried to boxes. For example, $Q^{(n)}$ denotes an element in a sequence of partitions and $B^{(n)}(x)$ is that box in $Q^{(n)}$ containing x.

Think of $(X_1, Y_1), (X_2, Y_2),...$ as being observed in sequence, and suppose that at each stage we apply a recursive partitioning algorithm to U. We then obtain a triangular array $Q^{(n,j)}$ of partitions of U such that: (i) for n fixed, $Q^{(n,j+1)}$ is a refinement of $Q^{(n,j)}$ and (ii) each partition is composed of a set of boxes. Typically there is no refinement relation between $Q^{(n,j)}$ and $Q^{(n',j')}$ for $n \neq n'$. Proposition 3.6 is addressed to that issue.

3. The Principal Theorem

The following simple lemma, whose proof is left to the reader, is the vital observation connecting our previous paper [6] and the present one.

3.1. Lemma. *Let (X, Y) be a random vector for which X takes values in U, the unit cube in R^d, and Y takes values in R^1; $E\{|Y|\} < \infty$. Let F be the marginal distribution of X, and for Borel $A \subset U$ let $G\{A\} = E\{YI_{[X \in A]}\}$, where $I_{[X \in A]}$ is the indicator of $\{X \in A\}$. Define $h(X)$ to be $E\{Y \mid X\}$. Conclude that $h(X) = (dG/dF)(X)$.*

Before we can proceed with our results we need to establish some further notation and a definition. In what follows if $B^{(n)}(x)$ is a member of the partition $Q^{(n)}$, and F is as in Lemma 3.1, then we write $F(y \mid B^{(n)}(x)) = P\{Y \leqslant y \mid X \in B^{(n)}(x)\}$ if $F\{B^{(n)}(x)\} > 0$.

3.2. Definition. If $Q^{(n)}$ is a partition of U and F and G are as in Lemma 3.1, then if $Y \geqslant 0$, we denote a particular version of $(dG/dF)(X)$ with respect to $Q^{(n)}$ by $h_n(X)$, where

$$\begin{aligned} h_n(x) &= \int_0^\infty (1 - F(y \mid B^{(n)}(x))\, dy \qquad \text{if} \quad F\{B^{(n)}(x)\} > 0 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{3.3}$$

Notice that, in a slight abuse of notation, $h_n(X)$ is a version of $E\{Y \mid Q^{(n)}\} = E\{h(X) \mid Q^{(n)}\}$. The extension of (3.3) to the case of general Y is clear from the decomposition $Y = Y^+ - Y^-$, where Y^+ and Y^- are nonnegative. The next lemma will permit us to conclude that the convergence

in pth mean of these conditional expectations follows from their convergence in probability.

3.4. LEMMA. *If the assumptions of Lemma* 3.1 *hold, and* $E\{|Y|^p\} < \infty$ *for some* $p \geqslant 1$, *then the sequence* $\{|h_n(X)|^p\}$ *is uniformly integrable.*

Proof. Without loss suppose Y, and therefore $h_n(X)$, is nonnegative; the extension to the general case is straightforward. Now $(h_n(X))^p \leqslant E\{Y^p \mid Q^{(n)}\}$ almost surely according to Jensen's inequality. Set $V_{n,k} = \{E\{Y^p \mid Q^{(n)}\} > k\}$. Then

$$P(V_{n,k}) \leqslant \frac{E\{E\{Y^p \mid Q^{(n)}\}\}}{k} = \frac{E\{Y^p\}}{k}$$

according to the Markov inequality; it follows that $P(V_{n,k})$ tends to 0 uniformly in n as k tends to ∞. From the definition of conditional expectation

$$E\{E\{Y^p \mid Q^{(n)}\} I_{V_{n,k}}\} = E\{Y^p I_{V_{n,k}}\}.$$

The latter is not more than the

$$\sup_{A:P(A) \leqslant E\{Y^p\}/k} E\{Y^p I_A\},$$

which tends to 0 as k tends to ∞ since $E\{Y^p\} < \infty$.

The next result is a restatement of Proposition 2.10 of [6] adapted to the purpose of this paper. The proof of that proposition was incomplete. With our new formulation we offer here a proof which is both brief and complete.

3.5. DEFINITION. If $Q^{(n)}$ is a partition of U, $x \in U$, and $\operatorname{supp} X$ is the support of F, then

$$D_n(x) = \max_i \sup\{|z_i - y_i| : y, z \in B^{(n)}(x) \cap \operatorname{supp} X\}.$$

3.6. PROPOSITION. *Let* (X, Y), F, G, *and* h *be as in Lemma* 3.1. *Let* $Q^{(n)}$ *be a sequence of partitions of* U *which satisfies*

$$D_n(X) \text{ tends in probability to } 0 \text{ as } n \text{ tends to } \infty. \tag{3.7}$$

Suppose further that for fixed $p \geqslant 1$, $E\{|Y|^p\} < \infty$. *It follows that* $E\{|h_n(X) - h(X)|^p\}$ *tends to* 0 *as* n *tends to* ∞.

Proof. In view of Lemma 3.4 our conclusion obtains if we show that $h_n(X) - h(X)$ tends to 0 in probability.

Let $\varepsilon > 0$ be arbitrary and g be a bounded, continuous function on U for which $E\{|h(X) - g(X)|\} < \varepsilon$. Write $g_n(X) = E\{g(X) \mid Q^{(n)}\}$. Jensen's

inequality for conditional expectations implies that $E\{|h_n(X) - g_n(X)|\} < \varepsilon$. Because g is a bounded, continuous function on a compact set, it is uniformly continuous. Therefore, there exists $\delta > 0$ such that if $x \in \operatorname{supp} X$ and $D_n(x) < \delta$, then $|g_n(x) - g(x)| < \varepsilon$. Since $P\{D_n(X) < \delta\}$ tends to 1 as n tends to ∞, our proposition is proven.

We now show that under conditions like those of (2.32) of [6] that a certain class of truncated means tend in pth mean to the regression function $h(X)$ they are trying to estimate, provided only that $h(X)$ has a finite absolute pth moment. The algorithms to which these results apply all involve

$$\begin{aligned} &\text{sequences of successively refined,} \\ &\text{possibly data dependent, partitions:} \\ &Q^{(n,1)},\ldots, Q^{(n,n)} = Q^{(n)} \text{ of } U \text{ as in (3.07a) of [6].} \end{aligned} \tag{3.8}$$

Moreover, we require

$$\begin{aligned} &\text{sequences } k(n) \text{ and } \gamma(n),\ n = 1, 2,\ldots \text{ for which} \\ &k(n)/n \to 0,\ \gamma(n) \to \infty, \text{ and } \gamma(n)\sqrt{n}/k(n) \to 0 \text{ as } n \to \infty \\ &\text{(in particular, } k(n)/\sqrt{n} \to \infty \text{ as } n \to \infty). \end{aligned} \tag{3.9}$$

For any given random sample $(X_1, Y_1),\ldots, (X_n, Y_n)$ each distributed as (X, Y), $Q^{(n)}$, $k(n)$, and $\gamma(n)$, our estimator $\hat{h}_n(X)$ of $h(X)$ is defined to be

$$\begin{aligned} \hat{h}_n(X) = \sum_{i=1}^{n} &[-\gamma(n) \vee (\gamma(n) \wedge Y_i)]\, I_{[X_i \in B^{(n)}(X)]}/n\hat{F}_n\{B^{(n)}(X)\} \\ &\text{if} \quad \hat{F}_n\{B^{(n)}(X)\} > k(n)/n \\ = 0 \quad &\text{otherwise.} \end{aligned} \tag{3.10}$$

Here for $A \subset U$, $\hat{F}_n\{A\}$ is the empirical probability of A based on $X_1,\ldots, X_n$.

In many cases $\gamma(n)$ grows so fast that $\gamma(n) \geqslant \max_{1 \leqslant i \leqslant n} |Y_i|$; in that case $\hat{h}_n$ is an untruncated local average for $\hat{F}_n\{B^{(n)}(X)\}$ large enough. For example, if the common distribution of the Y_i's is normal, and $\gamma(n)$ grows as fast as some power of n, then almost surely ultimately $\hat{h}_n$ is an untruncated local average.

It would make sense to define $\hat{h}_n$ to be $\bar{Y}$ or some estimate of $E\{Y\}$ other than 0 when $\hat{F}_n\{B^{(n)}(X)\} \leqslant k(n)/n$. Our proofs would necessarily then be more complicated. More important, for the algorithms we study the difference between the two $\hat{h}_n$'s tends to 0 in L^p whenever Y has a finite pth moment.

The next two definitions (one of which was introduced in [6]) are important because with them and the theorems which involve them we are able to verify that (3.7) holds for all (X, Y), F, and G for certain algorithms. In what follows, distribution functions are assumed to be right continuous.

3.11. DEFINITIONS. If H is a Borel probability on U with ith coordinate marginal distribution H_i, and Q is a partition of U composed of boxes, then the ith *norm of Q relative to H* is

$$\|Q\|_i^H = \sum \{[H_i(b(B)) - H_i(a(B))]\, H\{B\} \mid B \in Q\};$$

the ith *norm of Q relative to $H-$* is

$$\|Q\|_i^{H-} = \sum \{[H_i(b(B)-) - H_i(a(B)-)]\, H\{B\} \mid B \in Q\}.$$

3.12. THEOREM. *If $\|Q^{(n)}\|_i^{\hat{F}_n}$ and $\|Q^{(n)}\|_i^{\hat{F}_n-}$ both tend to 0 in probability, (3.8) and (3.9) hold, and $\hat{F}_n\{x \mid \hat{F}_n(B^{(n)}(x)) > k(n)/n\}$ tends to 1 in probability, then $D_n(X)$ tends to 0 in probability.*

Notice that norms relative to $\hat{F}_n$ can be computed from data.

Proof of the Theorem. From two applications of the argument at the bottom of page 527 of [6] it follows that

$$\|Q^{(n)}\|_i^{\hat{F}_n} - \|Q^{(n)}\|_i^F \qquad \text{and} \qquad \|Q^{(n)}\|_i^{\hat{F}_n-} - \|Q^{(n)}\|_i^{F-}$$

both tend in probability to 0. Therefore, both $\|Q^{(n)}\|_i^F$ and $\|Q^{(n)}\|_i^{F-}$ tend in probability to 0.

Choose K with $F\{K\} = 1$ so that if $x \in K$ (with ith coordinate x_i), then for every i either x_i is an atom of F_i or for every $d > 0$ both $F_i(x_i + d) - F_i(x_i)$ and $F_i(x_i) - F_i(x_i - d)$ are positive.

The Markov inequality implies that for every $\varepsilon > 0$, $F\{x \mid F_i(b_i(B^{(n)}(x))) - F_i(a_i(B^{(n)}(x))) > \varepsilon\}$ tends in probability to 0. If $x \in K$ and x_i is not an atom of F_i, then $F_i(b_i(B^{(n)}(x))) - F_i(a_i(B^{(n)}(x)))$ tends to 0 only if $b_i(B^{(n)}(x)) - a_i(B^{(n)}(x))$ does.

Suppose now that x_i is an atom of F_i. Clearly $F_i(b_i(B^{(n)}(x))) - F_i(a_i(B^{(n)}(x))) \geqslant F_i\{x_i\}$ if $a_i(B^{(n)}(x)) < x_i$. It follows that $I_{\{a_i(B^{(n)}(x)) = x_i\}}$ tends in probability to 1. A similar argument with $F-$ shows that $I_{\{b_i(B^{(n)}(x)) = x_i\}}$ tends in probability to 1. Fix $\gamma > 0$ and small. Let $K_\gamma \subset K$ satisfy: $F\{K_\gamma\} > 1 - \gamma$ and the projection of K_γ on the ith coordinate axis contains only finitely many atoms of F_i. From what has been shown it follows that $\sup\{|z_i - y_i| \mid y, z \in B^{(n)}(X) \cap K_\gamma\}$ tends in probability to 0. But γ was arbitrary. Therefore, $\sup\{|z_i - y_i| \mid y, z \in B^{(n)}(X) \cap K\}$ tends in probability to 0. And the latter is almost surely equal to $\sup\{|z_i - y_i| \mid y, z \in B^{(n)}(X) \cap \operatorname{supp} X\}$. Since $i \in \{1,..., d\}$ was arbitrary, our theorem is proven.

If Y^+ (Y^-) is the positive (negative) part of Y, then, as has been mentioned, the representation $Y = Y^+ - Y^-$ entails that $h(X)$, as defined previously, can be represented as a difference of two nonnegative random variables, $h = h^+ - h^-$. Likewise, $\hat{h}_n$ can be decomposed into such a difference, say $\hat{h}_n = \hat{h}_n^+ - \hat{h}_n^-$. If we can show that $\hat{h}_n^+$ tends in pth mean to

$\hat{h}^+$, then the same argument applies as well to $\hat{h}_n^-$, and we are done. Of course, $Q^{(n)}$ is generated from the full sample $(X_1, Y_1),..., (X_n, Y_n)$. In view of the foregoing discussion it should be clear that we may conveniently suppose that we observe a "training sample" $V_1,..., V_n$, where $V_i = (X_i, Y_i, Z_i)$, and a component X of (X, Y, Z), and we wish to estimate $h(X) = E\{Y \mid X\}$; the partition $Q^{(n)}$ is determined by $V_1,..., V_n$. While we have suggested that Z is the negative part of Y, in the previous notation, we do not preclude the case in which the Z_i's include information from, for example, extraneous randomization.

3.13. THEOREM. *Let $V, V_1,..., V_n$ be iid, $V = (X, Y, Z)$, $X \in R^d$, $Y, Z \in (R^1)^+$; $E(Y^p) < \infty$ for some $p \geqslant 1$. Suppose* (3.8) *and* (3.9) *hold, and that $\hat{h}_n$ is defined as in* (3.10). *Suppose further that*

$$\hat{F}_n\{x \mid \#_n(B^{(n)}(x)) > k(n)\} \tag{3.14}$$

tends in probability to 1, *where $\#_n(B)$ is the number of $X_1,..., X_n$ in the box B and that*

$$\|Q^{(n)}\|_i^{\hat{F}_n} \text{ and } \|Q^{(n)}\|_i^{\hat{F}_n -} \text{ tend in probability to } 0. \tag{3.15}$$

It follows that $\hat{h}_n$ tends in p-th mean to h.

Proof. We prove only the case $E\{Y\} = 1$, leaving the trivial extension to the general case to the reader. Notice that (3.10) can ultimately be rewritten in an obvious notation as

$$\begin{aligned} \hat{h}_n(X) &= \int_0^{\gamma(n)} \{1 - \hat{F}_n(y \mid B^{(n)}(X))\}\, dy \qquad \text{if} \quad \#_n(B^{(n)}(X)) > k(n). \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{3.16}$$

Write

$$h_n(X) = \int_0^{\infty} \{1 - F(y \mid B^{(n)}(X))\}\, dy \tag{3.17}$$

and recall that h_n is the Radon–Nikodym derivative of G with respect to F when both measures are restricted to $Q^{(n)}$.

If we show that $\hat{h}_n(X) - h_n(X)$ tends in pth mean to 0, then we may apply Proposition 3.6 and Theorem 3.12 to complete the proof. Write

$$K_n = \sup_{x,y} |\hat{F}_n(x, y) - F(x, y)|,$$

where $F(x, y)$ is the joint cumulative distribution function of (X, Y). From Kiefer [7], $P(\sqrt{n}\, K_n > r) < C \exp\{-r^2\}$, where C depends only on the

dimension d. It follows that $\sqrt{n}\,K_n$ is bounded in pth mean, uniformly in F and n. Moreover, whenever $\#_n(B^{(n)}(X))$ exceeds $k(n)$

$$|\hat{h}_n(X) - h_n(X)| \leqslant \int_0^{\gamma(n)} |\{1 - F(y \mid B^{(n)}(X))\} - \{1 - \hat{F}_n(y \mid B^{(n)}(X))\}|\, dy$$

$$+ \int_{\gamma(n)}^{\infty} \{1 - F(y \mid B^{(n)}(X))\}\, dy = \mathrm{I} + \mathrm{II}.$$

Now

$$\mathrm{II} = E\{(Y - \gamma(n))\, I_{\{Y > \gamma(n)\}} \mid Q^{(n)}\}(X),$$

which tends to 0 in pth mean as a consequence of Jensen's inequality and the finiteness of $E\{Y^p\}$.

It is helpful to denote the numerator of $F(y \mid B^{(n)}(X))$ by $P_n(y, X)$ and the numerator of $\hat{F}_n(y \mid B^{(n)}(X))$ by $\hat{P}_n(y, X)$. Then

$$I = \int_0^{\gamma(n)} |F(y \mid B^{(n)}(X)) - \hat{F}_n(y \mid B^{(n)}(X)|\, dy$$

$$= \int_0^{\gamma(n)} \left| \frac{\hat{P}_n(y, X)}{\hat{F}_n\{B^{(n)}(X)\}} - \frac{P_n(y, X)}{F\{B^{(n)}(X)\}} \right| dy$$

$$= \int_0^{\gamma(n)} \left| \frac{\hat{P}_n(y, X)}{\hat{F}_n\{B^{(n)}(X)\}} - \frac{P_n(y, X)}{\hat{F}_n\{B^{(n)}(X)\}} + \frac{P_n(y, X)}{\hat{F}_n\{B^{(n)}(X)\}} - \frac{P_n(y, X)}{F\{B^{(n)}(X)\}} \right| dy$$

$$\leqslant \frac{1}{\hat{F}_n\{B^{(n)}(X)\}} \int_0^{\gamma(n)} |\hat{P}_n(y, X) - P_n(y, X)|\, dy$$

$$+ \int_0^{\gamma(n)} P_n(y, X) \left| \frac{1}{F_n\{B^{(n)}(X)\}} - \frac{1}{\hat{F}_n\{B^{(n)}(X)\}} \right| dy = \mathrm{III} + \mathrm{IV}.$$

Recall that $\hat{F}_n\{B^{(n)}(X)\} \geqslant k(n)/n$. Also, $|\hat{P}_n(y, X) - P_n(y, X)| \leqslant (d + 1)\, 2^d K_n$. Therefore, $\mathrm{III} \leqslant (d + 1)\, 2^d (n/k(n))\, \gamma(n)\, K_n = (d + 1)\, 2^d (\sqrt{n}\, \gamma(n)/k(n))(\sqrt{n}\, K_n)$. Now $\sqrt{n}\, K_n$ is bounded in pth mean, and thus (3.9) implies that III tends in L^p to 0. Also,

$$\mathrm{IV} = \int_0^{\gamma(n)} P_n(y, X) \left| \frac{F\{B^{(n)}(X)\} - \hat{F}_n\{B^{(n)}(X)\}}{F\{B^{(n)}(X)\}\, \hat{F}_n\{B^{(n)}(X)\}} \right| dy$$

$$= \int_0^{\gamma(n)} F(y \mid B^{(n)}(X)) \left| \frac{F\{B^{(n)}(X)\} - \hat{F}_n\{B^{(n)}(X)\}}{\hat{F}_n\{B^{(n)}(X)\}} \right| dy$$

$$\leqslant (d+1)\, 2^d K_n \frac{n}{k(n)} \int_0^{\gamma(n)} F(y \mid B^{(n)}(X))\, dy \leqslant (d+1)\, 2^d \frac{\gamma(n)\sqrt{n}}{k(n)} (\sqrt{n}\, K_n),$$

which tends in L^p to 0.

The proof of Theorem 3.13 is now complete.

Notice that Theorem 3.13 holds for $\gamma(n)$ which depends on $V_1,\ldots, V_n$ provided only that $\gamma(n)$ tends in probability to ∞ and that $\gamma(n)K_n/(k(n)/n)$ tends in pth mean to 0.

4. Quantile Cuts

The condition (3.15) is essential to Theorem 3.13 because it permits the application of Theorem 3.12 at a crucial juncture. The purpose of this section is to recall how (3.15) can be guaranteed for an algorithm. We draw upon [6] and the notion of a quantile cut.

For a given box, B, we say informally that an ith *p-quantile cut* has been achieved if the box is refined by a cut perpendicular to coordinate axis i so that at most p of the original contents of B lies in either of its two daughter boxes. Necessarily p is at least 1/2. Of course, for a preassigned box B and number p between 1/2 and 1 it may not be possible to perform an ith p-quantile cut when the marginal distributions of F are not continuous. In that case a more technical notion is necessary, and the interested reader is referred to Sections 3 and 4 of [6] for the precise definitions. In any case, the following lemma, whose proof follows from the proof of (3.09) of [6], connects quantile cuts and (3.15).

4.1. Lemma. *Suppose there exist monotone nondecreasing sequences* $m_n \to \infty$ *and* $1/2 \leqslant q_n \leqslant 1$ *for which* $q_n^{m_n}$ *tends to* 0 *and* $\hat{F}_n\{x \mid$ *for at least* m_n *indices* j, $B^{(n,j+1)}(a(B^{(n,j)}(x))$ *and* $B^{(n,j+1)}(b(B^{(n,j)}(x))$ *comprise an* i*-th* q_n*-quantile cut of* $B^{(n,j)}(x)$ *relative to* $\hat{F}_n\}$ *tends to* 1 *in probability for each coordinate axis* i. *It follows that* (3.15) *holds.*

5. Untruncated Local Averages and the Relationship to Stone's Work

In this our concluding section we revert to the scenario of Section 1: $(X, Y), (X_1, Y_1), (X_2, Y_2),\ldots$ iid, $X \in R^d$, $Y \in R^1$. The set $\{(X_i, Y_i)\}$ is called a "training sample of size n." We sketch here how within box untruncated averages can be studied from the same point of view as was taken by Stone in his masterful study of nearest-neighbor estimators of regression functions [10]. The reader is urged to have Stone's paper and [6] at hand when reading this section.

Stone's estimators of $E\{Y \mid X\}$ are weighted sums of the Y_i's in the training sample. His weights depend only on X and on the X_i's of the training sample. For a sequence of weights which are nonnegative and sum to 1 his

conditions for L^p consistency have three parts, which can be described roughly as follows. The estimators must be asymptotically local. Thus, if X_i is far from X, then Y_i should, in a sense he makes precise, figure little in the estimation of $E\{Y \mid X\}$. Also, the weights should be asymptotically negligible, so that no fixed, finite number of members of the training sample should have large influence on the estimator. Finally, the estimator of any nonnegative function of X should not have expectation larger than a fixed multiple of the expectation of the function itself. It is the last condition cited—condition (1) of Stone's Theorem 1—which is typically the most difficult to verify. Also, his mathematical arguments do not apply immediately to our recursive partitioning estimators because when the latter are rewritten in his form, the weights depend on the Y_i's of the training sample as well as on X and the X_i's.

We begin our rigorous discussion with the case in which Y has a finite range: $\{y_1,..., y_r\}$. By analogy to the classification problem it is helpful to think of (X_i, Y_i), $i = 1,..., n$, sampled as follows. First, Y_i is chosen, with $\pi_j = P(Y_i = y_j)$. Then, given that $Y_i = y_j$, X_i is drawn from F_j, the conditional distribution of X_i given that $Y_i = y_j$. An algorithm for generating the partitions $Q^{(n)}$ is assumed given. Another X, independent of $\{(X_i, Y_i)\}$ is observed; its corresponding Y is not observed. We want $E\{Y \mid X\}$, and, in a notation consistent with Stone, we denote our estimate of $E\{Y \mid X\}$ by $\hat{E}_n = \hat{E}_n\{Y \mid X\}$, which we write as follows:

$$\hat{E}_n\{Y \mid X\} = \sum_{j=1}^{r} y_j \hat{\pi}_{j,n} \frac{\hat{F}_{j,n}\{B^{(n)}(X)\}}{\hat{F}_n\{B^{(n)}(X)\}}, \tag{5.1}$$

where $\hat{F}_{j,n}$ is the empirical distribution of those X_i among $\{X_1,..., X_n\}$ for which $Y_i = y_j$ and $\hat{\pi}_{j,n}$ is the fraction of $\{Y_1,..., Y_n\}$ which equal y_j. If $\hat{\pi}_{j,n} = 0$, we take $\hat{F}_{j,n} \equiv 0$. With this definition of $\hat{E}_n$, (2.32) of [6] as completed in Section 3, is easily brought to bear on consistency results. A comparison of our work and that of Stone is made easier if $\hat{E}_n$ is rewritten:

$$\hat{E}_n\{Y \mid X\} = \sum_{i=1}^{n} W_{ni}(X, X_1,..., X_n, Y_1,..., Y_n)\, Y_i, \tag{5.2}$$

where

$$\begin{aligned} W_{ni} = W_{ni}(X, X_1,..., X_n, Y_1,..., Y_n) & \\ &= [\#_n(B^{(n)}(X))]^{-1} \quad \text{if} \quad X_i \in B^{(n)}(X) \\ &= 0 \qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{5.3}$$

Note that W_{ni} as written in (5.3) depends on $Y_1,..., Y_n$ through the rule for partitioning which determines $B^{(n)}(X)$. As has been noted, Stone's W's

depend only on the X's. This dependence renders his Fubini argument ([10, p. 611]) for his Theorem 1 invalid for our $\hat{E}_n$. Regardless, when Y has finite range, as we have assumed for the present, from (2.32) of [6] and Slutsky's theorem it follows that our $\hat{E}_n\{Y \mid X\}$ tends to $E\{Y \mid X\}$ in probability as n tends to infinity, provided the conditions of (2.32) of [6] are satisfied. (These are the relevant conditions of Theorem 3.13 of the present paper.) Because Y is bounded, convergence is in mean of order p for every p.

We need to extend the foregoing to a general distribution for Y. Thus, suppose Y has an arbitrary distribution for which $E\{|Y|^p\} < \infty$ for some $p \geqslant 1$, but that our assumptions are otherwise as they were. Refer to the paragraph of Stone's paper [10, p. 611] which begins "Consider." There is no loss in replacing his $Y^{(M)}$ there by a random variable (which we also denote by $Y^{(M)}$) which not only is bounded but also has finite range. Now $\hat{E}_n\{Y^{(M)} \mid X\}$ tends in L^p to $E\{Y^{(M)} \mid X\}$ if the conditions of (2.32) of [6] are satisfied. Clearly, $\hat{E}_n\{Y \mid X\}$ will tend in L^p to $E\{Y \mid X\}$ provided also that

$$\lim_{M \to \infty} E\left\{\left|\sum_{i=1}^{n} W_{ni}(Y_i - Y_i^{(M)})\right|^p\right\} = 0 \qquad \text{uniformly in } n. \tag{5.4}$$

The conditions of (2.32) of [6] do not suffice to guarantee that (5.4) holds. However, from the arguments in the cited paragraph of [10] it follows that the condition (1) of Stone's Theorem 1 is sufficient to guarantee (5.4):

> There exists $0 < C < \infty$ such that for every nonnegative Borel f on R^d,
>
> $$E\left\{\sum_{i=1}^{n} W_{ni} f(X_i)\right\} \leqslant CE\{f(X)\}. \tag{5.5}$$

The next step in our discussion is to state a condition which guarantees that (5.5) holds. To that end, define

$$\begin{aligned} U_{ni} = U_{ni}(X, X_1, \ldots, X_i, \ldots, X_n, Y_1, \ldots, Y_i, \ldots, Y_n) \text{ to be} \\ W_{ni}(X_i, X_1, \ldots, X, \ldots, X_n, Y_1, \ldots, Y, \ldots, Y_n). \end{aligned} \tag{5.6}$$

Stone's argument for his Proposition 11 ([10, p. 613]) applied to our W_{ni} and U_{ni} shows that (5.5) is implied by the existence of a universal bound to $\sum_{i=1}^{n} U_{ni}$.

We can summarize the preceding four paragraphs as follows. The sequence of $\hat{E}_n$ of (5.1) and (5.2) tends to $E\{Y \mid X\}$ in mean of order p whenever (1) $E\{|Y|^p\} < \infty$, (2) the conditions of (2.32) of [6] are satisfied, and (3) $\sum_{i=1}^{n} U_{ni}$ is bounded by a universal constant. Our difficulties in providing partitions which satisfy (3) were what led us to the approach of Section (3).

Anderson rules [1] can be approached from the point of view of this section. Most of our results for these rules require that

$$F \text{ has continuous coordinatewise marginal distributions,} \tag{5.7}$$

and

$$\text{there exists } M > 1 \text{ such that for each } Q^{(n)},\ k(n) \geqslant 3 \text{ and } k(n) \leqslant \#_n(B^{(n)}(x)) \leqslant Mk(n) \text{ for all } x, \text{ where } k(n)/\sqrt{n} \to \infty \text{ and } k(n)/n \to 0 \text{ as } n \to \infty. \tag{5.8}$$

Before we define the rules we need to define the concept of local rank.

5.9. DEFINITION. For a given $\{x_1, x_2, ..., x_n\}$, where $x_k \in U$ for each k, and a box $B \subset U$, the ith *local rank* of x_j among $\{x_1, ..., x_n\}$ evaluated at the box B, written $r_{B,i}^{(n)}(x_j)$, is the rank of the ith coordinate of x_j among the ith coordinates of those $x_k \in B$.

Notice that for $r_{B,i}^{(n)}(x_j)$ to require no further specification, for each (n, i) there can be no ties among the ith coordinates of those $x_k \in B$. When (5.7) holds, and $\{x_1, ..., x_n\}$ are realizations of iid vectors each distributed as F, then almost surely all local ranks are defined.

An Anderson rule (which produces "statistically equivalent blocks") can be described completely by a unique sequence of binary "decision" trees (one for each n) and pairs of numbers attached to each node. The first of these numbers indicates the axis i on which the box associated with that node is to be cut; the second denotes the ith local rank of the observation at which the cut is to be made. (Implicit in this representation is some convention such as requiring that the endpoint always be assigned to the left daughter node.) Strictly speaking, an "Anderson rule" is a sequence of rules, one for each n, but for convenience we refer to both an individual rule and the entire sequence as "the" rule. In what follows we use the symbol $\mathcal{T}$ to denote a binary tree; which one will be clear from context.

For Anderson rules upper bounds on the functions $\|Q^{(n)}\|_i^{\hat{F}_n}$ and $\|Q^{(n)}\|_i^{\hat{F}_n -}$ can be forced to tend to 0, since (when (5.7) holds) quantile cuts can always be implemented in such a way that the cuts depend solely on local ranks. The argument for Lemma (3.09) of [6] applies here; the reader should note that the discussion surrounding (3.16) of [6] is not relevant. It follows that (3.15) and therefore the relevant conditions of (2.32) of [6] can be guaranteed. We will show that there are Anderson rules for which $\sum_{i=1}^{n} U_{ni}$ is bounded by a universal constant. (The reader will note that Anderson rules do not depend on $Y_1, ..., Y_n$, so for them the foregoing extensions of Stone's arguments are not required.)

We make careful distinction in the discussion to follow between nodes and

boxes. (This distinction is like the distinction between a function and its value.) References to nodes are to corresponding paths down a tree and do not depend on a particular set of data. References to boxes must be qualified by the data which give rise to each specific box. For a given set of data, to each node of the tree $\mathscr{E}$ of an Anderson rule there is naturally associated a box in a partition of U.

We shall use the symbol $\mathscr{D}_n^*$ to refer to a set whose $n+1$ members belong to U; write $\mathscr{D}_n^* = \{x_0, x_1, ..., x_n\}$. For a given $\mathscr{D}_n^*$ write $\mathscr{D}_{0,n} = \{x_1, ..., x_n\}$, and for $i = 1, 2, ..., n$, $\mathscr{D}_{i,n} = \{x_1, ..., x_{i-1}, x_0, x_{i+1}, ..., x_n\}$. Occasionally the members of the various $\mathscr{D}$'s will be called "observations." We think of them as realizations of the corresponding X's.

Now let $\mathscr{D}_n^*$ be given. For $k = 0, 1, ..., n$ let $B_{(N,k)}^{\mathscr{E}}$ be a generic symbol for a box corresponding to node N of the decision tree of an Anderson rule with corresponding tree $\mathscr{E}$ when the data are $\mathscr{D}_{k,n}$. We shall make use of the following statement.

$\mathscr{P}_r$: If the decision tree $\mathscr{E}$ of an Anderson rule has no more than r total nodes, then for every node N, $B_{(N,0)}^{\mathscr{E}}$ and $B_{(N,i)}^{\mathscr{E}}$ have at least $\#_n(B_{(N,0)}) - 1 = \#_n(B_{(N,i)}) - 1$ points in common, and all the local ranks of their points in common differ by at most 1.

5.10. LEMMA. *If all local ranks are defined for $\mathscr{D}_{0,n}, ..., \mathscr{D}_{n,n}$, then $\mathscr{P}_r$ holds for all Anderson rules and all positive integers r (necessarily $r \leqslant 2n-1$).*

Proof. $\mathscr{P}_1$ is immediate since the corresponding rule does not partition the data. $\mathscr{P}_2$ is exactly $\mathscr{P}_1$ since there are no binary trees with two nodes. We now argue $\mathscr{P}_3$. Let N be the root node and N' and N'' be the daughter nodes. The local ranks relative to N differ by at most 1. Therefore, the same observations of $B_{(N,i)}^{\mathscr{E}}$ and $B_{(N,0)}^{\mathscr{E}}$ go into left and right daughters associated with N' and N'', respectively, save perhaps for that x whose ith coordinate is the cut point.

For $m \geqslant 3$ assume that $\mathscr{P}_m$ holds. In order to establish $\mathscr{P}_{m+1}$ we study an Anderson rule with $m+1$ total nodes and corresponding tree $\mathscr{E}$. An Anderson rule whose tree agrees with the tree under consideration save for one eliminated splitting at terminal node N has a corresponding tree $\mathscr{E}'$ to which $\mathscr{P}_m$ applies. For $\mathscr{D}_{0,n}(\mathscr{D}_{i,n})$ the passage from $\mathscr{E}'$ to $\mathscr{E}$ involves one split of $B_{(N,0)}^{\mathscr{E}'}(B_{(N,i)}^{\mathscr{E}'})$ in a preassigned direction and at an observation with preassigned local rank. There are two cases to consider. It may happen that $B_{(N,0)}^{\mathscr{E}'}$ and $B_{(N,i)}^{\mathscr{E}'}$ consist of identical observations. In that case the conclusion of $\mathscr{P}_{m+1}$ is immediate. Otherwise, there are exactly $\#_n(B_{(N,0)}^{\mathscr{E}'}) - 1 = \#_n(B_{(N,i)}^{\mathscr{E}'}) - 1$ observations in common to $B_{(N,0)}^{\mathscr{E}'}$ and $B_{(N,i)}^{\mathscr{E}'}$. The respective local ranks of the observations in common differ by at most one, and the argument for $\mathscr{P}_3$ shows that $\mathscr{P}_{m+1}$ holds.

If $\mathscr{D}_n^*$ and an Anderson rule with decision tree $\mathscr{T}$ are given, then for $x \in U$, $N(x \mid \mathscr{D}_{i,n})$ is the unique node N of $\mathscr{T}$ for which $x \in B_{(N,i)}^{\mathscr{T}}$.

5.11. DEFINITIONS. If $\mathscr{D}_n^*$ and $\mathscr{T}$ are as above, then the *invariant subset* $\mathscr{I}$ of $\mathscr{D}_{0,n}$ is $\{x_j \in \mathscr{D}_{0,n}: N(x_j \mid \mathscr{D}_{0,n}) = N(x_j \mid \mathscr{D}_{i,n})$ for $i = 1, 2, ..., n\}$. An observation $x_j \in \mathscr{D}_{0,n}$ which is not invariant is called *variant.*

5.12. LEMMA. *If an Anderson rule satisfies* (5.8) *and all local ranks are defined, then* $\mathscr{D}_{0,n} \backslash \mathscr{I}$ *has cardinality at most* $3n/k(n)$.

Proof. Since an Anderson rule nonadaptively determines an axis—call it d_0—and an d_0 local rank on which to cut a given $B_{(N,i)}^{\mathscr{T}}$, Lemma 5.10 implies that the only points in $B_{(N,i)}^{\mathscr{T}}$ which are variant because of a cut of that box are those within one d_0 local rank of the cut point in $B_{(N,0)}^{\mathscr{T}}$. There are three such points in $B_{(N,0)}^{\mathscr{T}}$. Because of (5.8) the tree corresponding to our Anderson rule has at most $n/k(n)$ terminal nodes. Moreover, in every binary tree there are one fewer nonterminal nodes than terminal nodes. Therefore, fewer than $n/k(n)$ $B_{(N,i)}^{\mathscr{T}}$ are partitioned by the rule when $\mathscr{D}_{i,n}$ are data. Hence, our lemma is established.

5.13. LEMMA. *Assume that* $\mathscr{D}_n^*$ *and an Anderson rule (with tree* $\mathscr{T}$*) are given. Suppose that the rule satisfies* (5.8) *and that all local ranks are defined. It follows that the cardinality of* $\{N(x_0 \mid \mathscr{D}_{i,n}) \mid i = 1, 2, ..., n\}$ *is at most* 2^d.

Proof. As x_0 traverses $\mathscr{T}$ from its root node, let N_1 be the first node N such that as x_0 passes through N its assignment to a daughter node of N is not identical for all $\mathscr{D}_{i,n}$. We call N_1 the "first ambiguous node." Of course, there may be no such node N_1. In that case the cardinality of $\{N(x_0 \mid \mathscr{D}_{i,n}) \mid i = 1, 2, ..., n\}$ is 1. Thus, suppose there exists N_1 and that d_1 is the axis on which a cut is prescribed at N_1.

Since we have assumed that depending on the choice of data $\mathscr{D}_{i,n}$ x_0 sometimes passes through the left daughter node of N_1 and sometimes the right, its local rank must sometimes exceed N_1's cutting threshold and sometimes not. By (5.10) the value of the local rank across all possible choices of data can change by at most two. In view of (5.8) the d_1st local ranks of x_0 in the daughter box and all subsequent daughter boxes is sufficiently small or sufficiently large so as to ensure that x_0 is never again within one d_1st local rank of an observation which is cut on the d_1st axis.

We can define two "second ambiguous nodes," N_2^L and N_2^R; $N_2^L(N_2^R)$ is the first ambiguous node for the subtree of $\mathscr{T}$ which has root the left (right) daughter node of N_1. If N_1 exists, but neither N_2^R nor N_2^L does, then the cardinality of $\{N(x_0 \mid \mathscr{D}_{i,n}) \mid i = 1, 2, ..., n\}$ is exactly 2. If all three exist, then the cardinality in question is at least 4. The notion of ambiguous node can

be defined successively, and any path from the root node of $\mathscr{E}$ to a terminal node encounters at most one ambiguous node at which is specified a cut on any fixed axis. Therefore, any path which traverses $\mathscr{E}$ from root node to a terminal node encounters at most d ambiguous nodes. In view of the foregoing, it follows from an induction on the maximal number of ambiguous nodes per path from the root of $\mathscr{E}$ to a terminal node that the cardinality of $\{N(x_0 \mid \mathscr{D}_{i,n}) \mid i = 1, 2, \ldots, n\}$ is at most 2^d.

In Lemmas 5.10, 5.12, and 5.13 we have implicitly understood Anderson rules to be operating on the range of the X's in their partitioning of the feature space U. We now revert to the scenario of the earlier part of this section and think of the rules as operating on the training sample and given X. The estimators $\hat{E}_n$ are, as before, given by (5.1)–(5.3), and the U_{ni} are given by (5.6).

5.14. LEMMA. *Suppose that an Anderson rule satisfies* (5.7) *and* (5.8), *and that* $\hat{E}_n$ *is given by* (5.1)–(5.3). *It follows that there is a universal bound which is almost surely not exceeded by* $\sum_{i=1}^{n} U_{ni}$.

Proof. Assumption (5.7) implies that almost surely all local ranks are defined. Lemma 5.13 entails that for each n, X lies in terminal boxes associated with at most $2d$ terminal nodes of the corresponding tree $\mathscr{E}$ when the roles of X and X_i ($i \leqslant n$) are interchanged (as in (5.6)). Therefore, for each n, X lies in terminal boxes under interchanges with at most $2dMk(n)$ invariant X_i's. Lemma 5.12 implies that for each n there are at most $3n/k(n)$ variant X_i's. It follows that

$$\sum_{i=1}^{n} U_{ni} \leqslant \frac{1}{k(n)} \{2dMk(n)\} + \frac{1}{k(n)} \left\{ \frac{3n}{k(n)} \right\}$$

$$\leqslant 2dM + \sup_n \frac{3n}{k^2(n)}.$$

The latter supremum is finite (from (5.8), and it clearly does not depend on n), so our lemma is proven.

Lemma 5.14 and the arguments which preceded (5.7) can be combined into the following result.

5.15. THEOREM. *Assume that* (5.7) *holds and that for a given Anderson rule* (5.8) *holds. Moreover, suppose that* $\|Q^{(n)}\|_i^{\hat{E}_n}$ *and* $\|Q^{(n)}\|_i^{\hat{E}_n -}$ *tend to* 0 *for each* i. *It follows that* $\hat{E}_n$ *tends to* $E\{Y \mid X\}$ *in mean of order* p *for every* $p \geqslant 1$ *for which* $E\{|Y|^p\} < \infty$.

ACKNOWLEDGMENT

We gratefully acknowledge the considerable help of Charles Stone with our work on this paper.

REFERENCES

[1] ANDERSON, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. In *Multivariate Analysis* (P. R. Krishnaiah, Ed.), pp. 5–27. Academic Press, New York.

[2] BREIMAN, L., FRIEDMAN, J. H., RAFSKY, L., AND STONE, C. J. (1980). Growing and pruning trees for classification and regression, to appear.

[3] COX, D. R. (1970). *The Analysis of Binary Data*. Methuen, London.

[4] DEVROYE, L. P., AND WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

[5] FRIEDMAN, J. H. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Computers* **C-26** 404–408.

[6] GORDON, L., AND OLSHEN, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Ann. Statist.* **6** 515–533.

[7] KIEFER, J. (1961). On large deviations of the empiric d.f. of vector chance variables and a law of iterated logarithm. *Pacific J. Math.* **11** 649–660.

[8] SONQUIST, J. (1970). *Multivariate Model Building: The Validation of a Search Strategy*. Institute for Social Research, Univ. of Michigan, Ann Arbor.

[9] SPIEGELMAN, C., AND SACKS, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* **8** 240–246.

[10] STONE, C. J. (1977). Nonparametric regression and its applications (with discussion). *Ann. Statist.* **5** 595–645.

[11] SUTHERLAND, D. H., OLSHEN, R. A., COOPER, L., AND WOO, S. L. (1980). The development of mature gait. *J. Bone Joint Surgery* **62A** 336–353.